Jimin Sun

+1(412) 694-2980 ◇ Pittsburgh, PA
jimins2@cs.cmu.edu ◇ jiminsun.github.io ◇ github.com/jiminsun

INTERESTS Natural Language Processing, Multilingual NLP, Domain Adaptation, Language Pragmatics

EDUCATION

Carnegie Mellon University — Pittsburgh, PA
Master of Language Technologies, Language Technologies Institute (Advisors: Yonatan Bisk, Jean Oh) — Aug 2023

Seoul National University — Seoul, South Korea
M.S., Industrial Engineering (GPA: 3.85/4.0) — Aug 2021

Carnegie Mellon University — Pittsburgh, PA
Visiting student funded by Ministry of Science and ICT (GPA: 3.93/4.0) — Feb 2021

Seoul National University — Seoul, South Korea
B.A., Psychology (Minor in Statistics) (GPA: 3.59/4.0) — Aug 2018

EXPERIENCE

Cohere AI — Toronto, Canada (Remote)
Machine Learning Intern — May 2022 - Aug 2022
- Built a multilingual toxicity detection system in five languages (en, de, fr, es, pt) for an internal toxicity filter to ensure safe text generation of large language models

Kakao Enterprise — Pangyo, South Korea
AI Research Intern — Feb 2021 - Jul 2021
- Developed methods to adapt machine translation models to chat & medical domain, validating their effectiveness at WMT21
- Designed pipeline to crowd-source English-Korean translations, control data quality, collect human judgments for evaluation

PUBLICATIONS

Jimin Sun, Patrick Fernandes, Xinyi Wang, Graham Neubig, "A Multi-dimensional Evaluation of Tokenizer-free Multilingual Pretrained Models" (2022), https://arxiv.org/abs/2210.07111

Jimin Sun, Ye Won Byun, Cathy Jiao, Shahriar Noroozizadeh, Rosa Vitiello, "ET tu, CLIP? Addressing Common Object Errors for Unseen Environments" (2022), Conference of Computer Vision and Pattern Recognition - Embodied AI Workshop. https://embodied-ai.org/papers/2022/20.pdf

Yunju Bak, Jimin Sun, Jay Kim, Sungwon Lyu, Changmin Lee, "Kakao Enterprise's WMT21 Machine Translation using Terminologies Task Submission" (2021), Proceedings of the Sixth Conference on Machine Translation (WMT). https://aclanthology.org/2021.wmt-1.79/

Jimin Sun, Hwijeen Ahn, Chan Young Park, Yulia Tsvetkov, David R. Mortensen, "Cross-Cultural Similarity Features for Cross-Lingual Transfer Learning of Pragmatically Motivated Tasks" (2021), Conference of the European Chapter of the Association for Computational Linguistics (EACL). https://aclanthology.org/2021.eacl-main.204/

Jimin Sun, Hwijeen Ahn, Chan Young Park, and Jungyun Seo, "NLPDove at SemEval-2020 Task 12: Improving Offensive Language Detection with Cross-lingual Transfer" (2020), Proceedings of the Fourteenth Workshop on Semantic Evaluation, 1576–1586. https://www.aclweb.org/anthology/2020.semeval-1.206/

PROJECTS

A Multi-dimensional Evaluation of Tokenizer-free Multilingual Pre-trained Models — Jan 2022 - Jun 2022
- Independent Study Project for Spring 2022 submitted to EMNLP 2022 (Advisor: Graham Neubig)
- Performed a comprehensive empirical comparison of multilingual tokenizer-free and subword-based models considering fine-tuning robustness, data efficiency, inference cost
- Ran extensive experiments using five pretrained models for three tasks (Natural Language Inference, Named Entity Recognition, Extractive QA) covering over 20 languages

CVPR 2022 Embodied AI Workshop: Addressing Object Errors for Unseen Environments — Jan 2022 - May 2022
- CMU 11-777 (Multimodal ML) Semester Project accepted to CVPR 2022 Embodied AI Workshop (Advisor: Yonatan Bisk)
- Suggested a method to incorporate the pretrained CLIP model to the Episodic Transformer architecture for the ALFRED task
- Found that the suggested method helps with leveraging visual descriptions, detecting small objects, understanding rare words

Safe and Seamless Operation of Manned and Unmanned Aircraft in Shared Airspace Sep 2021 - May 2022

- Advisors: Jean Oh, Sebastian Scherer
- Developed Automatic Speech Recognition system to transcribe pilot conversations
- Incorporated Speech Generation models to enable the interaction of an AI pilot with human pilots

Pre-training Character-level Encoders for Morphologically-rich Languages Sep 2021 - Dec 2021

- CMU 11-711 (Advanced Natural Language Processing) Semester Project (Advisor: Graham Neubig)
- Suggested a pre-training objective suitable for morphologically rich languages (Kiswahili), which outperformed CANINE, a SoTA character-level encoder in question answering benchmark (TyDi QA)

Conference on Machine Translation (WMT21) Shared Task: MT using Terminologies May 2021 - July 2021

- Developed COVID-19 domain translation model in four language pairs based on transformers, improving performance using domain data selection, back-translation (Ranked 1st out of 22 submissions for English→French)
- Proposed new approach to enforce terminology constraints, improving keyword accuracy (+4.7% Exact Match Accuracy) while preserving general translation quality (+.5 BLEU)

SKT AI Fellowship: Korean News Summarization using KoBERT, KoGPT Jun 2020 - Nov 2020

- Built Korean news summarization model based on three architectures (Pointer-generator Network, Presumm, BART), investigating the benefits and caveats of leveraging pre-trained models for summarization tasks
- Won grant of $10,000 among 40 teams and achieved *Best Application Award* at final presentation

SemEval 2020: Improving Offensive Language Detection with Cross-lingual Transfer Dec 2019 - Apr 2020

- Attended SemEval-2020 (Offensive Language Detection), achieving competitive results in all five languages (1st place in Greek)
- Suggested cross-lingual transfer approach to find samples in high-resource languages that helps task in low-resource language

RELEVANT COURSES

- Introduction to Machine Learning (CMU 10-701); Advanced Natural Language Processing (CMU 11-711); Multimodal Machine Learning (CMU 11-777); Machine Learning for Text Mining (CMU 11-741); Algorithms for NLP (CMU 11-711); Artificial Intelligence: Representation and Problem Solving (CMU 15-381); IoT, Big Data, and Machine Learning (CMU 17-640); Deep Learning (SNU M2177.003100)

AWARDS

- Intensive AI Program at Carnegie Mellon University (scholarship sponsored by the South Korean government) 2019 - 2020
- SK Telecom Best Application Award (Grant of $10,000 among 40 teams) 2019
- Eminence scholarship, Seoul National University 2015
- Merit-based scholarship, Seoul National University 2013, 2014, 2015

SERVICE

- Point of contact for two potential incoming students during CMU LTI Open House 2022
- CMU LTI Application Support Program for two students, providing feedback on Statements of Purpose and CV 2021
- Hosted a department wide LTI seminar on Multilingual NLP with Prof. David Mortensen, inviting three guests in academia and industry 2021